The 19th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM 2025)

Can User Feedback Help Issue Detection? An Empirical Study on a One-billion-user Online Service System

Shuyao Jiang[1], Jiazhen Gu[1], Wujie Zheng[2,3], Yangfan Zhou[2], Michael R. Lyu[1]

1. Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China

2. College of Computer Science and Artificial Intelligence, Fudan University, Shanghai, China

3. Tencent Inc., Shenzhen, China

User Feedback and Issue Detection

- **User feedback** in large-scale online service systems
  - Describe user experience with the systems
  - An important data resource for service maintenance (e.g., issue detection)

- Challenges of user feedback for issue detection
  - User feedback is abundant but noisy
  - Hard to identify severe issues automatically

How user feedback help issue detection?

Research Questions

- **RQ1:** What **proportion** of feedback actually reports issues?

- **RQ2:** Does **feedback amount** indicate issue severity?

- **RQ3:** Can **certain features** (*e.g.*, sentiment, text length, historical behaviors) of a feedback item indicate issue severity?

- **RQ4:** Are **feedback topics** stable over time?

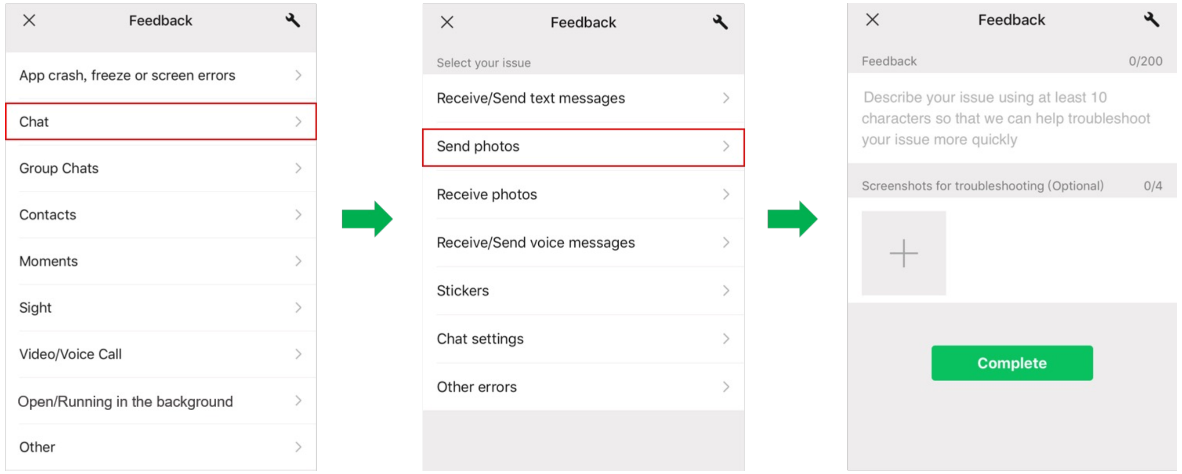
Target System and Dataset

- Target service system: **WeChat**
  - A popular social media platform serving over one billion users
  - Contains diverse online services (e.g., chat, work, entertainment)

- User feedback dataset
  - Six online services from WeChat ecosystem
  - All feedback collected within one year (a total of 50,378,766 items)

TABLE I
SIX SERVICES OF OUR TARGET SOCIAL MEDIA PLATFORM.

Service	Functionality	User Scale
WeChat	Instant Messaging	10^9
WeChat-Work	Workplace Communication	10^8
WeChat-info	Content Subscription	10^8
WeChat-Pay	Mobile Payment	10^8
WeChat-Game	Mobile Game	10^8
WeChat-Reading	Reading	10^8



RQ1: Proportion of Issue-relevant Feedback

- Step 1: Manual analysis
  - Randomly select 10,000 feedback items for manual analysis
  - A large amount (4,450) is *irrelevant* to system issues
- Step 2: Model training
  - Train a **binary classifier** using a labeled dataset of the 10,000 feedback items
    - BERT + TextCNN (F1 = 88.71%)
- Step 3: Prediction for the entire feedback dataset

TABLE II
TOTAL AMOUNT OF FEEDBACK ITEMS AND ISSUE-RELEVANT FEEDBACK ITEMS COLLECTED FROM SIX SERVICES IN ONE YEAR.

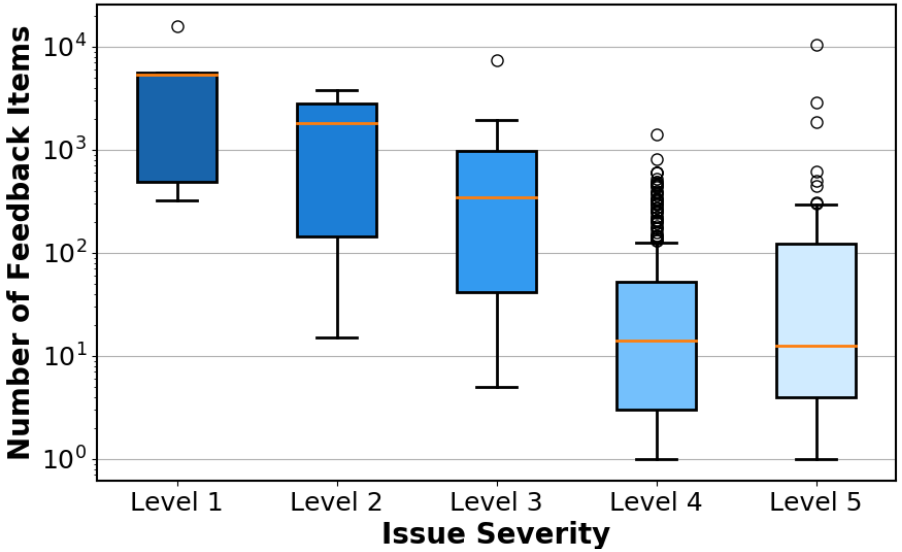
Service	**# Feedback**	**# Issue-relevant**	**Percentage**
WeChat	16,405,550	10,916,729	66.54%
WeChat-Work	20,406,139	2,232,664	10.94%
WeChat-Info	7,771,103	1,289,212	16.59%
WeChat-Pay	2,306,460	652,204	28.28%
WeChat-Game	2,897,146	1,752,791	60.50%
WeChat-Reading	592,368	174,179	29.40%

Answer to RQ1: Only 10.94% ~ 66.54% of user feedback was issue-relevant, revealing the necessity of automated filtering to remove noise.

RQ2: Feedback Amount vs. Issue Severity

- Does more feedback mean more severe issues?

- Analysis from the **issue-tracking system**
  - An **issue ticket** describes a known issue with its *severity level* (1~5, 1 is the most severe)
  - We analyzed all 509 issue tickets in one year and their related user feedback



Answer to RQ2: While severe issues generally attract more feedback, some critical issues were reported by a few users, limiting reliance on volume-based prioritization.

RQ3: Feedback Features vs. Issue Severity

Feature 1: **Sentiment**

Negative sentiment → ? → Severe issues

- Step 1: Sentiment analysis
  - Calculate a **sentiment score** for each feedback text
  - Groups: Negative / Neutral / Positive
- Step 2: Significance test (Z-test)
  - Is feedback in the **negative group** more likely to indicate severe issues?
    - If Z>1.65, yes

TABLE IV
PROPORTIONS OF FEEDBACK ITEMS INDICATING SEVERE ISSUES IN DIFFERENT SENTIMENT GROUPS AND THE Z-TEST RESULTS.

Service	**Sentiment**			**Z value**	
	Neg.	Neu.	Pos.	Neg.-Neu.	Neg.-Pos.
WeChat	24.2%	22.5%	19.6%	0.432	1.215
WeChat-Work	31.7%	24.6%	32.5%	1.308	-0.196
WeChat-Info	24.2%	20.8%	28.3%	0.874	-1.037
WeChat-Pay	19.6%	12.5%	14.7%	**2.114**	1.388
WeChat-Game	39.2%	28.8%	17.5%	**2.410**	**5.267**
WeChat-Reading	25.8%	23.3%	20.0%	0.636	1.520

* Neg.: Negative group, Neu.: Neutral group, Pos.: Positive group

The sentiment of feedback has no significant correlation with issue severity.

RQ3: Feedback Features vs. Issue Severity

Feature 2: **Text Length**

Long feedback texts → ? → Severe issues

- Step 1: Text length analysis
  - Count **the number of characters** for each feedback text
  - Groups: Short / Medium / Long
- Step 2: Significance test (Z-test)
  - Is feedback in the **long-text group** more likely to indicate severe issues?

TABLE V
PROPORTIONS OF FEEDBACK ITEMS INDICATING SEVERE ISSUES IN DIFFERENT TEXT-LENGTH GROUPS AND THE Z-TEST RESULTS.

Service	**Text-length**			**Z value**	
	S.	M.	L.	L.-M.	L.-S.
WeChat	20.6%	24.0%	26.8%	0.615	1.611
WeChat-Work	25.2%	45.1%	37.1%	-1.388	**2.755**
WeChat-Info	22.8%	27.8%	28.8%	0.083	1.319
WeChat-Pay	13.0%	25.0%	18.9%	-1.207	**1.727**
WeChat-Game	29.4%	25.8%	31.4%	1.155	0.462
WeChat-Reading	21.3%	21.5%	27.5%	1.376	1.543

* S.: Short-text group, M.: Medium-text group, L.: Long-text group

The text length of feedback has no significant correlation with issue severity.

RQ3: Feedback Features vs. Issue Severity

Feature 3: **User Historical Behaviors**

Users who reported severe issues before → ? → Report severe issues again

- Historical behavior analysis
  - Among 159 users who reported severe issues, 80 users have submitted multiple feedback
  - 9 users have reported multiple severe issues, averaging 3 issues per user

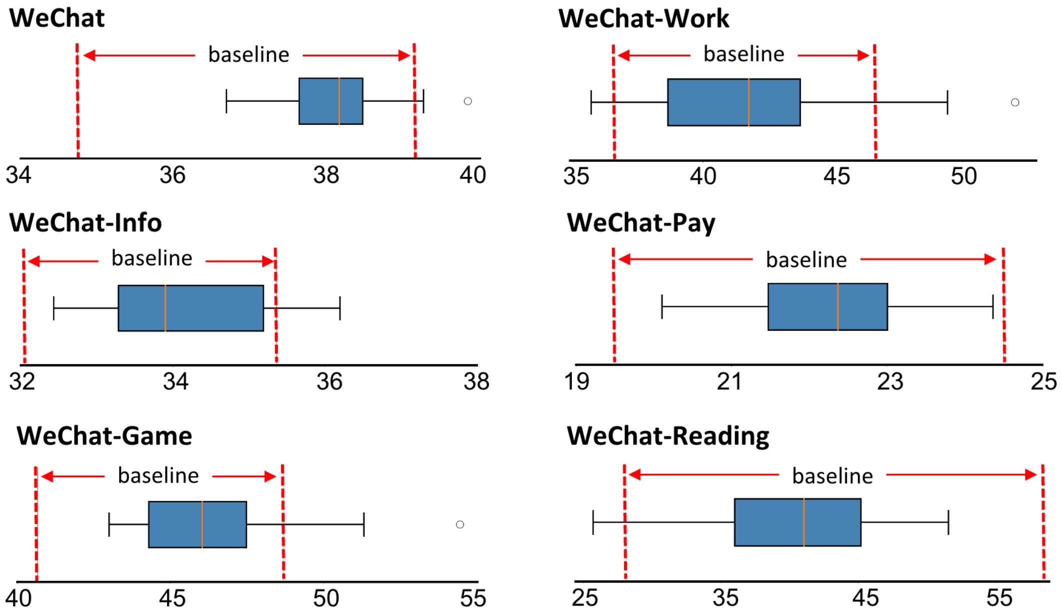
User historical behaviors can indicate issue severity to some extent.

Answer to RQ3: Text-based features (sentiment, text length) showed negligible correlation with issue severity, but historical user behavior (e.g., prior severe-issue reports) can offer some predictive value.

RQ4: Feedback Topic Stability Over Time

- Existing feedback analysis relies on **machine learning** methods
  - The stability of feedback topics affects the performance of AI-based analysis

- Feedback topic similarity analysis
  - Collect issue-relevant feedback from 8 service versions
  - Vectorize feedback texts
  - Measure the *Wasserstein Distance* between versions

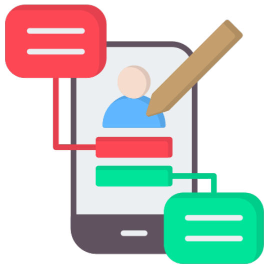


Answer to RQ4: Feedback topic distributions remained stable across service versions and time intervals, validating the feasibility of machine learning for longitudinal analysis.

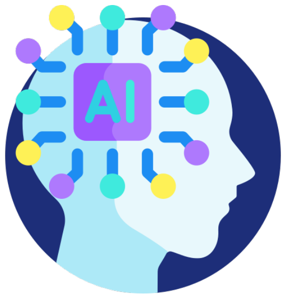
Summary of Key Findings

- High noise in feedback → Filtering needed
- Amount ≠ severity → Need for better prioritization
- Text features are poor predictors → Explore user behavior
- Topics are stable → AI-based analysis is feasible

Implications for Practice



- Improve **feedback interface** design
  - e.g., guided categorization



- Use **AI techniques** for filtering and classification
  - Apply LLMs for large-scale feedback learning
- Combine feedback with **system KPIs**
  - Multiple data sources for better issue detection

Shuyao Jiang (Ph.D. Candidate)

The Chinese University of Hong Kong

syjiang21@cse.cuhk.edu.hk

Pre-print

Homepage

Thank you!

ARISE Automated Reliable Intelligent Software Engineering

香港中文大學 The Chinese University of Hong Kong

復旦大學 FUDAN UNIVERSITY

Tencent 腾讯